Open Data as an Enabler for Development in Computational Linguistics

The Working Group for Open Data in Linguistics

Open Knowledge Foundation

Christian Chiarcos[1], Sebastian Hellmann[2], Sebastian Nordhoff[3], Philipp Cimiano[4], John McCrae[4], Jonas Brekle[2], Judith Eckle-Kohler[5], Iryna Gurevych[5,6], Silvana Hartmann[5], Michael Matuschek[5], Christian M. Meyer[5], Richard Littauer[7]

1 Information Science Institute
University of Southern California

2 Department of Computer Science
University of Leipzig

3 Department of Linguistics
Max Planck Institute for
Evolutionary Anthropology, Leipzig

4 Excellence Cluster Cognitive
Interaction Technology (CITEC)
University of Bielefeld

5 Ubiquitous Knowledge Processing
Lab (UKP-TUDA)
Technische Universität Darmstadt

6 German Institute for Educational Research
and Educational Information
Technische Universität Darmstadt

7 Computational Linguistics
Department
Saarland University

The Open Linguistics Working Group of the Open Knowledge Foundation (OWLG)

The Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN) is an initiative of experts from different fields concerned with linguistic data, including academic linguistics (e.g. typology, corpus linguistics), applied linguistics (e.g. computational linguistics, lexicography and language documentation), and NLP (e.g. from the Semantic Web community). The primary goals of the working group are 1) promoting the idea of open linguistic resources 2) the development of means for their representation, and 3) encouraging the exchange of ideas across different disciplines. Here, we focus on one particular aspect of our work, the promotion of linked data in linguistics. Other activities include the collection of use cases and the development of best practice guidelines for the publication of linguistic resources under open licenses, the documentation of workflows and the organization of workshops (e.g., LDL-2012, AG2 at the DGfS 2012).

At the moment, the Working Group assembles 67 people from 29 different organizations and 10 countries. Our group is relatively small, but continuously growing and sufficiently heterogeneous. It includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology, just to name a few, so, the ground for fruitful interdisciplinary discussions has been laid out.

Within the OWLG, a general consensus has been established that Semantic Web formalisms provide crucial advantages for the publication of linguistic resources. As shown below, all major types of data and metadata relevant to linguistic data collections (lexical-semantic resources, annotated corpora, metadata repositories and typological databases) can be represented by means of RDF and OWL; they are thus structurally interoperable (using RDF as representation formalism), and conceptually interoperable (with metadata and annotations are modeled in RDF, different resources can be directly linked to a single repository). The OWLG encourages the use of open licenses: For resources published under open licenses, an RDF representation yields the additional advantage that resources can be interlinked, and it is to be expected that an additional gain of information arises from the resulting network of resources. RDF is usually not the most appropriate format for every individual domain taken on its own; for linking data from different domains, however, it is the only viable option at present.

Nevertheless, the OWLG is not restricted to RDF as a representation formalism. For use cases that require interoperability only within a particular domain, other, and more efficient, but domain-specific representation formalisms may be the format of choice, e.g., XML standoff formats for annotated corpora, or LMF for lexical resources.

Open Data in Linguistics

Among the broad range of problems associated with linguistic resources, we identified four major classes of problems and challenges during our discussions that may be addressed by the OWLG. First, there is a great uncertainty with respect to legal questions of the creation and distribution of linguistic data; second, there are technical problems such as the choice of tools, representation formats and metadata standards for different types of linguistic annotation; third, we have not yet identified a point of reference for existing open linguistic resources; finally, there is the agitation challenge, i.e., how (and whether) we should convince our collaborators to release their data under open licenses.

Tim Berners-Lee and the W3C have recently proposed a 5 star rating system for data on the web. The first star is achieved by publishing data on the web (any format) under open licences. From this perspective Open Data Licences (http://www.opendefinition.org/) play a central role in building a foundation for a Linguistic Linked Data Web, which can be exploited for research in Computer Linguistics and Linguistics in general.

Towards a Linguistic Linked Open (LLOD) Data Cloud

The Linked Open Data (LOD) cloud represents the resulting set of resources. If published as Linked Data, linguistic resources represented in RDF can be linked with resources already available in the LOD cloud. At the moment, the LOD cloud already covers a number of lexical-semantic resources, including WordNet, YAGO, OpenCyc, and the DBpedia. Other types of linguistic resources, (linguistic corpora, typological data collections, linguistic terminology repositories) are not present in the LOD cloud at all.

One prospective goal of the OWLG can be seen in the development of a LOD (sub-)cloud of linguistic resources, the Linguistic Linked Open Data (LLOD) cloud, where linguistic resources (lexicalsemantic resources, corpora, metadata repositories) are not only provided in an interoperable way (using RDF), but also freely accessible (under an open license) and linked with each other (so that applications can combine information from different knowledge sources). In this article, we describe ongoing activities in the OWLG that will eventually lead to the creation of such a LLOD cloud.

Technical Background

Before coming to the description of the OWLG and its activities, we give a brief introduction of the technologies and terminological conventions applied throughout this article, in particular the notions of RDF, OWL/DL, and the concept of Linked Data.

The Resource Description Framework (RDF, Klyne et al., 2004) was originally invented to provide formal means to describe any resource, both offline (e.g. books in a library), and online (e.g. PDF documents in an electronic archive). The data structures provided by RDF were, however, so general that its use has extended far beyond its original application scenario. RDF is based on the notion of triples, consisting of a predicate that links a subject to an object. In other words, RDF formalizes relations between resources as edges in a directed labelled graph: Subjects are identified using globally unique URIs and can point to (via the predicate) another URI in the object part. Alternatively, triples can have simple strings in the object part that annotate the subject resource. At the moment, RDF represents the primary data structure of the Semantic Web and on this basis, a rich ecosystem of format extensions and technologies has evolved, including APIs, RDF databases (triple stores), the query language SPARQL, etc. Infrastructures for linguistic resources can benefit from these achievements and the relatively large and active community maintaining and improving technologies and representation formalisms.

RDF is based on globally unique and accessible URIs and it was specifically designed to establish links between such URIs (or resources). This is captured in the Linked Data paradigm (Berners-Lee, 2006) that postulates four rules:

1) Referred entities should be designated in a globally unambiguous way by URIs.

2) These URIs should be resolvable over HTTP.

3) Data should be represented by means of standards such as RDF.

4) A resource should include links to other resources.

With these rules, it is possible to follow links between existing resources to find other, related, data and exploit network effects.

Selected Resources for the Linguistic Linked Open Data Cloud

DBpedia: A General-Purpose Knowledge Base for the Semantic Web

DBpedia (Lehmann et al., 2009) is a community effort to extract structured information fromWikipedia and to make this information available on the Web. The main output of the DBpedia project is a data pool that (1) is widely used in academics as well as industrial environments, that (2) is curated by the community of Wikipedia and DBpedia editors, and that (3) has become a major crystallization point and a vital infrastructure for the Web of Data. DBpedia is one of the most prominent Linked Data examples and presently the largest hub in the Web of Linked Data. The extracted RDF knowledge from the English Wikipedia is published and interlinked according to the Linked Data principles and made available under the same license as Wikipedia (CC-BY-SA).

In its current version 3.7 DBpedia contains more than 3.64 million things, of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases. The DBpedia data set features labels and abstracts for 3.64 million things in up to 97 different languages; 2,724,000 links to images and 6,300,000 links to external web pages; 6,200,000 external links into other RDF datasets, and 740,000 Wikipedia categories. The dataset consists of 1 billion RDF triples out of which 385 million were extracted from the English edition of Wikipedia and roughly 665 million were extracted from other language editions and links to external datasets (Bizer, 2011).

Uby: A Network of Lexical-Semantic Resources

Uby (Gurevych at al., 2012) is a large integrated lexical resource developed at the Ubiquitous Knowledge Processing Lab, TU Darmstadt.

It currently contains interoperable versions of 8 open resources in two languages: English WordNet, Wiktionary, Wikipedia, FrameNet and VerbNet, German Wikipedia, Wiktionary and multilingual OmegaWiki. Uby also provides open sense alignments between a subset of these resources. Uby will be released by the end of March along with a Java-API and conversions tools licensed under the open Apache license. Within the 5 star rating system Uby falls in the 3 star category.

Uby is based on ISO-LMF (Francopoulo et al. 2009), rather than RDF. LMF establishes interoperability between linguistic resources, but its XML serialization does not require the use of globally unique identifiers (URIs). It is therefore not part of the cloud diagram. However, the extension of LMF to include URIs (Francopoulo 2007), and full-fledged RDF linearizations of LMF have been suggested, e.g., in the context of the Lexicon Model for Ontologies (*Lemon*) as described by McCrae et al. (2011).

Linking DBpedia and Lexical-Semantic Resources

A recent effort at the AKSW Leipzig is dedicated to the development of an DBpedia-based open-source framework to extract semantic lexical resources (a ontology about language use) from Wiktionary (http://downloads.dbpedia.org/wiktionary). The data currently includes language, part of speech, senses, definitions, synonyms, taxonomies and translations for each lexical word. Main focus is on flexibility (to the loose schema) and configurability (towards differing language-editions of Wiktionary). The configuration uses a XML encoding language-mappings and templates containing placeholders, thus enables the addition of languages without altering the source code. The extracted data can (due to its semantically richness) be automatically transformed into the Lemon model or simpler domain specific formats. By offering a Linked Data service, we hope to extend DBpedia's central role in the LOD infrastructure to the world of Open Linguistics.

MASC in POWLA: An Open Corpus as Linked Data

The Manually Annotated Sub-Corpus (MASC, Ide et al., 2010) is a corpus of 500K tokens of contemporary American English text drawn from the Open American National Corpus, written and spoken, and chosen from several genres (http://www.anc.org/MASC). The MASC project is committed to a fully open model of distribution, without restriction, for all data and annotations produced or contributed.

MASC comprises various layers of annotations, including parts-of-speech, nominal and verbal chunks, constituent syntax, annotations of WordNet senses, frame-semantic annotations, document structure, illocutionary structure and other layers of annotation. As a multi-layer corpus, MASC is distributed in the GrAF format (Ide and Suderman, 2007), an XML standoff format with all annotations of a document grouped together in a set of XML files pointing to the same piece of primary data.

XML standoff formats can be difficult to process, and therefore, RDF/OWL formalisations of multi-layer corpora have been suggested early (Burchardt et al. 2008). POWLA (Chiarcos 2012) represents such a formalism to represent linguistic corpora by means of semantic web formalisms, in particular, OWL/DL.

The idea underlying POWLA is to represent linguistic annotations by means of RDF, to employ OWL/DL to define data types and consistency constraints for these RDF data, and to adopt these data types and constraints from an existing representation formalism applied for the loss-less representation of arbitrary kinds of text-oriented linguistic annotation within a generic exchange format. Unlike Buchardt et al. (2008), this approach is not specific to a particular subset of annotation layers in one specific corpus, but generic: POWLA is another linearization of the PAULA data model, that is also underlying PAULA XML, an XML standoff format developed at SFB 632 “Information Structure” (Chiarcos et al., 2008). With POWLA as an OWL/DL linearization of the PAULA data model, all annotations currently covered by PAULA (i.e. any text-oriented linguistic annotation) can be represented as part of the Linguistic Linked Open Data cloud. A converter from GrAF to POWLA, applied to data from the MASC, can be found under http://purl.org/powla.

Glottolog/Langdoc: Language Classification and Bibliographical Database

Glottolog/Langdoc (Nordhoff and Hammarström 2011, http://glottolog.livingsources.org) provides access to 180k references to descriptive literature treating (mostly) lesser-known languages which are interlinked with a very detailed language classification. Glottolog/Langdoc is maintained by the Max-Planck-Institute for Evolutionary Anthropology Leipzig; due to restrictions inherited from the original bibliographies the data are free for non-commerical use only (CC-BY-NC).

The references are collated from 20 different bibliographies. For standard bibliographical data such as author and title, Glottolog/Langdoc uses DCMI and BIBO. Additional information includes document type, language, and geographical region.

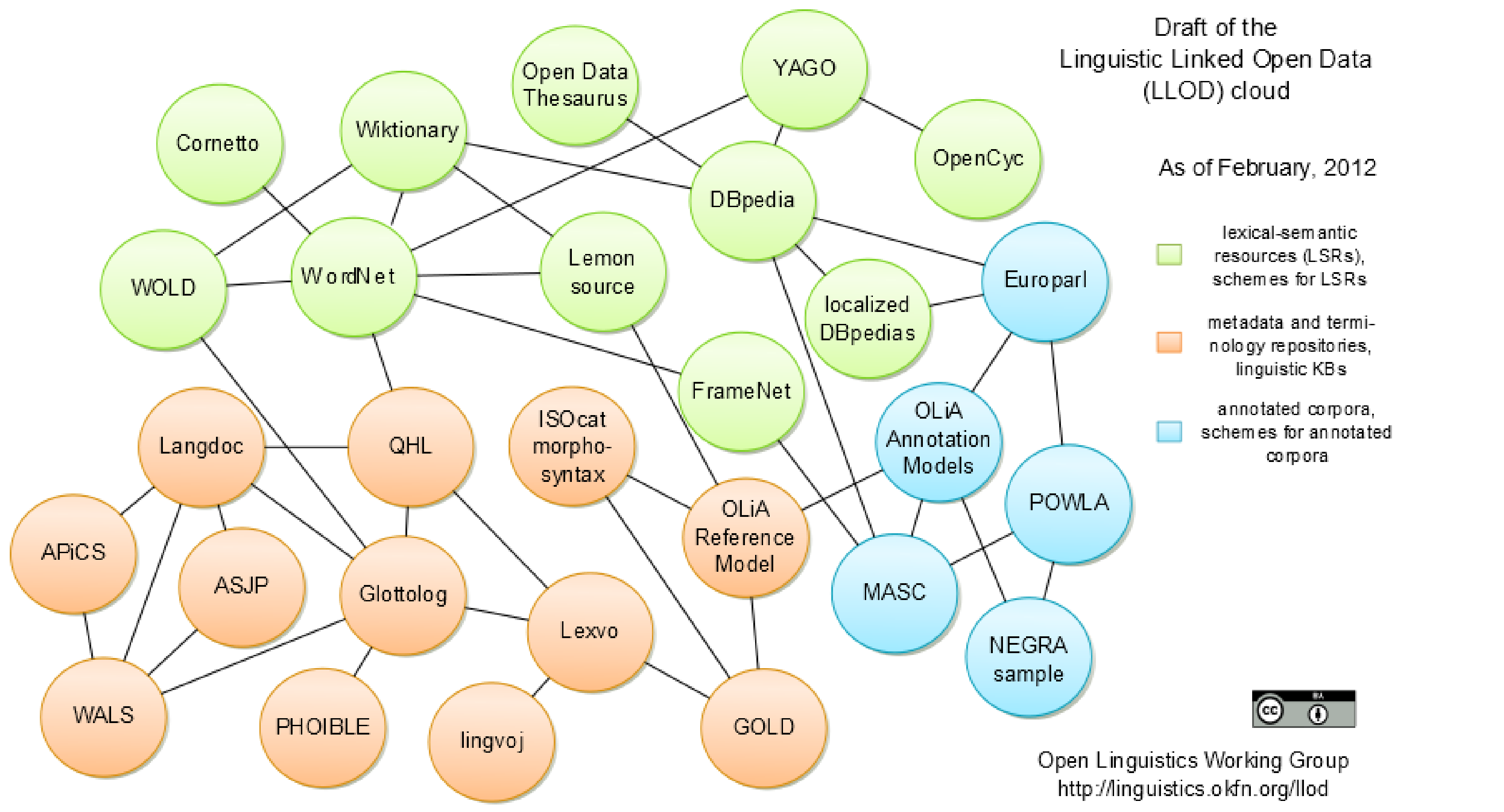
The bibliographical part Langdoc is complemented by the genealogical database Glottolog which lists names, codes, location, and family relations for 21288 languoids (languages, dialects, families), as well as a justification for why this languoid was included. Links to OLAC, Ethnologue, etc. are provided wherever possible. For Glottolog, a special purpose ontology had to be developed since ISO 639-3 based ontologies are unable to represent the required granularity..

Combining a bibliography with a genealogical database allows queries such as `Give me dictionaries of Afro-Asiatic languages from Africa'. All languoids and all references have their own URIs, allowing easy integration with other LLOD resources.

Other Resources

Aside from the resources mentioned here, the diagram includes further lexical-semantic resources, annotated corpora and linguistic data bases.

For details on these and the on-going development of the LLOD cloud, please consult http://linguistics.okfn.org/re-sources/llod.



Draft of the
Linguistic Linked Open Data
(LLOD) cloud

As of February, 2012

Open Linguistics Working Group
http://linguistics.okfn.org/llod

Linked Open Data as an Enabler for Development in Computational Linguistics

Representation and Modelling

Lexical-semantic resources can be described as labeled directed graphs (feature structures, Ide et al., 1995), as can annotated corpora (Bird and Liberman, 2001). RDF is based on labeled directed graphs and thus particularly well-suited for modeling both types of language resources.

Structural Interoperability

Using a common data model eases the integration of different resources: Merging multiple RDF documents yields another valid RDF document, while this is not necessarily the case for other formats. Moreover, HTTP allows multiple formats for the same resource to be published at the same location.

An important achievement as compared to domain-specific standards for resource interoperability (e.g., LMF for lexical-semantic resources, or GrAF for annotated corpora) is that RDF allows us to formulate queries that combine information from both sources, e.g., using the WordNet senses of the semantic annotation of the MASC corpus:

```
PREFIX wn20: <http://www.w3.org/2006/03/wn/wn20/schema/> .
PREFIX rkbWN: <http://wordnet.rkbexplorer.com/id/> .
SELECT ?token {
   rkbWN:synset-land-noun-2 wn20:containsWordSense ?sense .
   ?sense rdfs:label ?synonym .
   ?token powla:hasString ?synonym .
}
```

Semantic Interoperability

In a Linked Data approach, globally unique identifiers for concepts or categories can be used to define the vocabulary that we use and these URIs can be used by many parties who have the same interpretation of the concept. Furthermore, linking by OWL axioms allows to define the exact relation between two different concepts beyond simple equivalence statements.

Ecosystem and Technical Infrastructure

Linked data is supported by a community of developers in other fields beyond linguistics, and the ability to reuse their results is a clear advantage. One example is the Web Ontology Language OWL (McGuinness et al., 2004) that supports the formulation of axioms that constrain the way how the vocabulary is used, thus introducing the possibility of checking a lexicon or annotated corpus for consistency.

Federation and Querying Distributed Resources

In contrast to traditional methods, where it may be difficult to query across even multiple parts of the same resource, linked data allows for federated querying across multiple, distributed databases maintained by different data providers.

```
SELECT ?token {
   service <http://wordnet.rkbexplorer.com/sparql> {
      rkbWN:synset-land-noun-2
                                   wn20:containsWordSense
                                                   ?sense .
      ?sense rdfs:label ?synonym .
   }
   ?token powla:hasString ?synonym .
}
```

Dynamic Import

URIs can be used to refer external resources, we can thus import other linguistic resources “dynamically”: We can use URIs to point to other resources, they can be resolved when needed and will always contain the most recent version of the dynamically imported resources.

T. Berners-Lee, Timm (2006), Design Issues: Linked Data, http://www.w3.org/DesignIssues/LinkedData.html.

S. Bird and M. Liberman (2001). A formal framework for linguistic annotation. Speech Communication, 33(1):23–60.

C. Bizer (2011), DBpedia 3.7 released, including 15 localized editions. http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions.

A. Burchardt, S. Padó, D. Spohr, A. Frank, U. Heid (2008) Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In: Proceedings of the 3rd International Joint Conference on NLP (IJCNLP 2008), Hyderabad, India.

C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. TAL (Traitement automatique des langues), 49(2).

C. Chiarcos (2012), POWLA: Modeling linguistic corpora in OWL/DL. In: Proceedings of ESWC 2012, Heraklion, Greece, May 2012.

G. Francopoulo (2007), Strategy for an OWL specification of LMF, unpublished ms., Oct 2007, http://www.tagmatica.fr/lmf/StrategyForLMFInOWL29october2007.pdf.

G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet and C. Soria (2009), Multilingual resources for NLP in the lexical markup framework (LMF), Language Resources and Evaluation 43(1):57-70.

I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C.M. Meyer and C.Wirth (2012), Uby - A Large-Scale Unified Lexical-Semantic Resource, In: Proceedings of EACL 2012, Avignon, France, April 2012.

N. Ide, C. Fellbaum, C. Baker, and R. Passonneau (2010). The manually annotated subcorpus: A community resource for and by the people. In Proceedings of ACL-2010, pages 68–73.

N. Ide, J. Le Maitre and J. Véronis (1995). Outline of a model for lexical databases. In A. Zampolli, N. Calzolari, and M.S. Palmer, editors, Current Issues in Computational Linguistics: In Honour of Don Walker, pages 283–320. Giardini, Pisa, 1995.

N. Ide and K. Suderman (2007). GrAF: A graph-based format for linguistic annotations. In Proceedings of Linguistic Annotation Workshop (LAW 2007), pages 1–8.

G. Klyne, J.J. Carroll and B. McBride (2004), Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210.

J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann (2009). DBpedia - A crystallization point for the web of data. Journal of Web Semantics 7(3):154–165

J. McCrae, D. Spohr, and P. Cimiano (2011). Linking lexical resources and ontologies on the semantic web with Lemon. The Semantic Web: Research and Applications, pages 245–259.

D.L. McGuinness, F. van Harmelen and others (2004), OWL web ontology language overview, W3C recommendation, http://www.w3.org/TR/owl-features.

S. Nordhoff and H. Hammarström (2011), Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources, In: Proceedings of ISWC 2011, Bonn, Germany, Oct 2011.

printed by Universitätsrechenzentrum Leipzig